BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59233
doi: 10.3897/biss.4.59233

OPEN ACCESS

Conference Abstract

Unity in Variety: Developing a collection description standard by consensus

Matt Woodburn[‡], Deborah L Paul[§], Wouter Addink[|,¶,#], Steven J Baskauf[¤], Stanley Blum[«], Cat Chapman[»], Sharon Grant[^], Quentin Groom[˘], Janeen Jones[^], Mareike Petersen[¦], Niels Raes[?,|], David Smith[‡], Laura Tilley[ˀ], Maarten Trekels[˘], Michael Trizna[¢], William Ulate[ℓ,ꝏ], Sarah Vincent[‡], Ramona Walls[P,₳], Kate Webbink[^], Paula Zermoglio[₴]

‡ Natural History Museum, London, United Kingdom
§ Florida State University, Tallahassee, United States of America
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
# Species 2000 Secretariat, Leiden, Netherlands
¤ Vanderbilt University Libraries, Nashville, Tennessee, United States of America
« Biodiversity Information Standards (TDWG), San Francisco, United States of America
» University of Florida, Gainesville, United States of America
^ Field Museum, Chicago, United States of America
˘ Meise Botanic Garden, Meise, Belgium
¦ Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
? NLBIF, Leiden, Netherlands
ˀ CETAF, Brussels, Belgium
¢ Smithsonian Institution, Washington, United States of America
ℓ CRBio, San José, Costa Rica
ꝏ Missouri Botanical Garden, Saint Louis, United States of America
P University of Arizona, Tucson, United States of America
₳ CyVerse, Tucson, United States of America
₴ VertNet, Buenos Aires, Argentina

Corresponding author: Matt Woodburn (m.woodburn@nhm.ac.uk)

Received: 01 Oct 2020 | Published: 09 Oct 2020

Citation: Woodburn M, Paul DL, Addink W, Baskauf SJ, Blum S, Chapman C, Grant S, Groom Q, Jones J, Petersen M, Raes N, Smith D, Tilley L, Trekels M, Trizna M, Ulate W, Vincent S, Walls R, Webbink K, Zermoglio P (2020) Unity in Variety: Developing a collection description standard by consensus. Biodiversity Information Science and Standards 4: e59233. https://doi.org/10.3897/biss.4.59233

Abstract

Digitisation and publication of museum specimen data is happening worldwide, but far from complete. Museums can start by sharing what they know about their holdings at a higher level, long before each object has its own record. Information about what is held in collections worldwide is needed by many stakeholders including collections managers,

This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

funders, researchers, policy-makers, industry, and educators. To aggregate this information from collections, the data need to be standardised (Johnston and Robinson 2002).

So, the Biodiversity Information Standards (TDWG) Collection Descriptions (CD) Task Group is developing a data standard for describing collections, which gives the ability to provide:

1. automated metrics, using standardised collection descriptions and/or data derived from specimen datasets (e.g., counts of specimens) and
2. a global registry of physical collections (i.e., digitised or non-digitised).

Outputs will include a data model to underpin the new standard, and guidance and reference implementations for the practical use of the standard in institutional and collaborative data infrastructures.

The Task Group employs a community-driven approach to standard development. With international participation, workshops at the Natural History Museum (London 2019) and the MOBILISE workshop (Warsaw 2020) allowed over 50 people to contribute this work. Our group organized online "barbecues" (BBQs) so that many more could contribute to standard definitions and address data model design challenges. Cloud-based tools (e.g., GitHub, Google Sheets) are used to organise and publish the group's work and make it easy to participate. A Wikibase instance is also used to test and demonstrate the model using real data.

There are a range of global, regional, and national initiatives interested in the standard (see Task Group charter). Some, like GRSciColl (now at the Global Biodiversity Information Facility (GBIF)), Index Herbariorum (IH), and the iDigBio US Collections List are existing catalogues. Others, including the Consortium of European Taxonomic Facilities (CETAF) and the Distributed System of Scientific Collections (DiSSCo), include collection descriptions as a key part of their near-term development plans. As part of the EU-funded SYNTHESYS+ project, GBIF organized a virtual workshop: Advancing the Catalogue of the World's Natural History Collections to get international input for such a resource that would use this CD standard.

Some major complexities present themselves in designing a standardised approach to represent collection descriptions data. It is not the first time that the natural science collections community has tried to address them (see the TDWG Natural Collections Description standard). Beyond natural sciences, the library community in particular gave thought to this (Heaney 2001, Johnston and Robinson 2002), noting significant difficulties. One hurdle is that collections may be broken down into different degrees of granularity according to different criteria, and may also overlap so that a single object can be represented in more than one collection description. Managing statistics such as numbers of objects is complex due to data gaps and variable degrees of certainty about collection contents. It also takes considerable effort from collections staff to generate structured data about their undigitised holdings. We need to support simple, high-level collection summaries as well as detailed quantitative data, and to be able to update as needed. We

need a simple approach, but one that can also handle the complexities of data, scope, and social needs, for digitised and undigitised collections.

The data standard itself is a defined set of classes and properties that can be used to represent groups of collection objects and their associated information. These incorporate common characteristics ('dimensions') by which we want to describe, group and break down our collections, metrics for quantifying those collections, and properties such as persistent identifiers for tracking collections and managing their digital counterparts. Existing terms from other standards (e.g. Darwin Core, ABCD) are re-used if possible.

The data model (Fig. 1) underpinning the standard defines the relationships between those different classes, and ensures that the structure as well as the content are comparable across different datasets. It centres around the core concept of an 'object group', representing a set of physical objects that is defined by one or more dimensions (e.g., taxonomy and geographic origin), and linked to other entities such as the holding institution. To the object group, quantitative data about its contents are attached (e.g. counts of objects or taxa), along with more qualitative information describing the contents of the group as a whole.

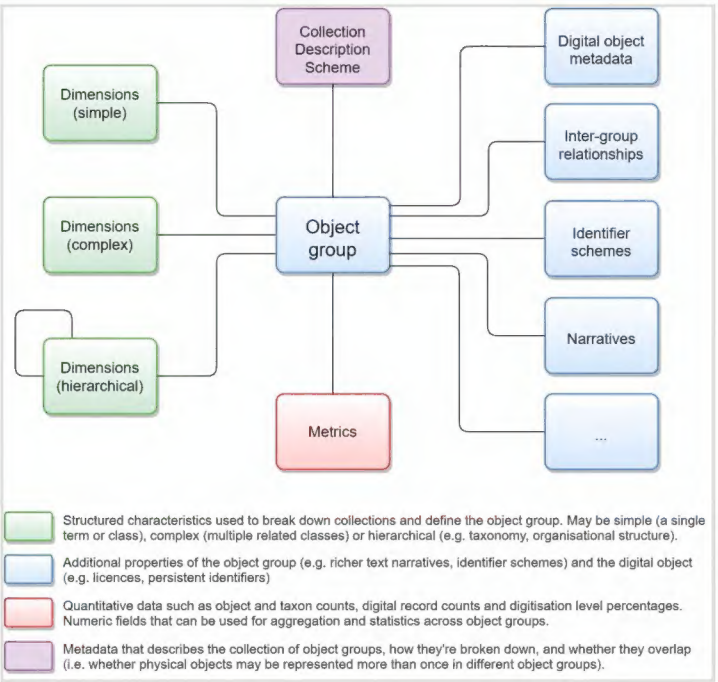

Figure 1.

A simplified representation of the TDWG CD data model

In this presentation, we will describe the draft standard and data model with examples of early adoption for real-world and example data. We will also discuss the vision of how the new standard may be adopted and its potential impact on collection discoverability across the collections community.

Keywords

collection descriptions, TDWG, data standards, biodiversity, geodiversity, natural sciences

Presenting author

Matt Woodburn

Presented at

TDWG 2020

Acknowledgements

Many thanks to all the interest and task group members contributing to this work.

Funding program

Support from COST (European Cooperation in Science and Technology) as part of the Mobilise Action CA17106 on Mobilising Data, Experts and Policies in Scientific Collections; and SYNTHESYS+ a Research and Innovation action funded under H2020-EU.1.4.1.2. Grant agreement ID: 823827.

References

- Heaney M (2001) An analytical model of collections and their catalogues. URL: https://ora.ox.ac.uk/objects/uuid:43a021d7-3024-4fd1-bca0-86028ec6ec7d
- Johnston P, Robinson B (2002) Collections and Collection Description. URL: http://www.ukoln.ac.uk/cd-focus/briefings/bp1/bp1.pdf